

# Pathway Fingerprinting - clustering tissue-specific data

Gabriel Altschuler

August 3, 2012

This document constructs consensus fingerprints for each of the tissue types profiled in the fingerprint paper and plots a heatmap clustering the different subtypes, as shown in Altschuler et al manuscript figure 2 and supplementary figure 4. A list of tissue-specific datasets was taken from *McCall et al. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res. (2011) 39(Database issue):D1011-5*. This is composed of arrays from 6 tissues; muscle, lung, spleen, kidney, liver and brain. The metadata is contained in a text file. The corresponding Pathway Fingerprints are obtained from the `pathprint` package.

```
> library(pathprint)
> # load data
> tissueData<-read.delim("tissueData.txt", stringsAsFactors = FALSE)
> # subset for data present in the fingerprint matrix
> tissueData<-tissueData[tissueData$GSM %in% GEO.metadata.matrix$GSM,]
```

Now a consensus pathway fingerprint can be created for each tissue/platform combination. A consensus threshold of 0.9 means that a pathways must have a score of +1 (or -1) in 90% of arrays within a set to have a score of +1 (or -1) in the consensus fingerprint. Plotting a heatmap shows the pathway fingerprint vectors and organizes the samples by hierachical clustering.

```
> tissueData$Type<-paste(tissueData$Tissue, tissueData$Platform, sep = "_")
> types<-unique(tissueData$Type)
> typeConsensus<-sapply(types, function(x){
  consensusFingerprint(GEO.fingerprint.matrix[,
    tissueData$GSM[tissueData$Type == x]],
    threshold = 0.9)
})

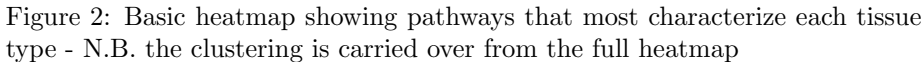
> # plot heatmap that contains only the pathways that vary across the dataset
> typeConsensus.SD<-apply(typeConsensus, 1, sd)
> full<-heatmap(typeConsensus[typeConsensus.SD > 0,], mar = c(10,10), scale = "none",
  keep.dendro = TRUE, col = c("blue", "white", "red"),
  cexCol = 0.75, cexRow = 0.5)
```



Figure 1 shows that the samples group by tissue type, rather than by platform (Altschuler et al figure 2 and supplementary figure 4). The pathways that most characterise each tissue and each tissue/platform combination can be highlighted, as shown in Altschuler et al figure 2.

```
> tissues<-unique(tissueData$Tissue)
> tissuesConsensus<-sapply(tissues, function(x){
  consensusFingerprint(GEO.fingerprint.matrix[,
    tissueData$GSM[tissueData$Tissue == x]],
    threshold = 0.99)
})
> # identify pathways unique to each tissue or tissue/platform combination
> tissuesConsensus.unique<-apply(tissuesConsensus, 1, function(x){min(table(x)) == 1})
> typeConsensus.unique<-apply(typeConsensus, 1, function(x){min(table(x)) %in% c(2,3,4)})

> # plot the heatmap
> heatmap(typeConsensus[
  (tissuesConsensus.unique & typeConsensus.unique),]
  , mar = c(10,20), scale = "none",
  Colv = full$Colv, col = c("blue", "white", "red"),
  cexCol = 0.75, cexRow = 0.7)
```



The R package @pheatmap@ was used to create the final figures for the manuscript. For the summary figure the very long pathway names were edited for space constraints and only the Wikipathways and Reactome pathways most significantly contributing to the clustering are displayed.

```
> library(pheatmap)
> names<-rownames(typeConsensus)
> names[grep("Respiratory electron transport", names)]<-
  "Respiratory electron transport (Reactome)"
> matrix<-typeConsensus[
  (tissuesConsensus.unique & typeConsensus.unique),]
> rownames(matrix)<-names[
  (tissuesConsensus.unique & typeConsensus.unique)]

> pheatmap(matrix, col = c("blue", "white", "red"),
  cexCol = 0.75, cexRow = 0.7, fontsize = 8,
  cellwidth = 5, cellheight = 5, legend = FALSE
)

> pheatmap(matrix[c(grep("Wikipathways", rownames(matrix)),
  grep("Reactome", rownames(matrix))),
  ],
  col = c("blue", "white", "red"),
  cexCol = 0.75, cexRow = 0.7, fontsize = 8,
  cellwidth = 10, cellheight = 10, legend = FALSE
)

> pheatmap(typeConsensus,
  # mar = c(10,20), scale = "none",
  #Colv = full$Colv,
  col = c("blue", "white", "red"),
  fontsize = 10,
  cellwidth = 10, #cellheight = 10,
  legend = FALSE, show_rownames = F,
  border_color = NA
)
```



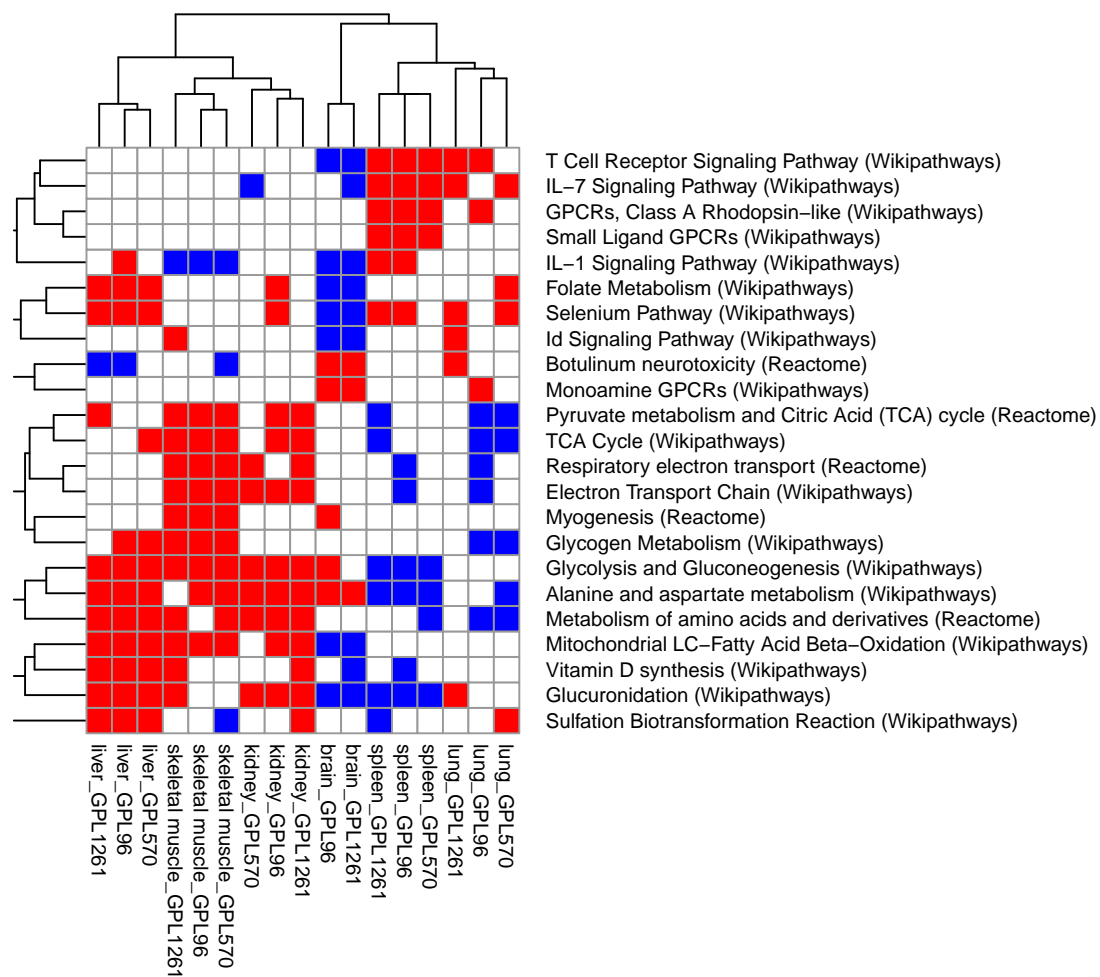


Figure 4: Heatmap formatted for the manuscript supplementary figure 4 - reduced pathway set

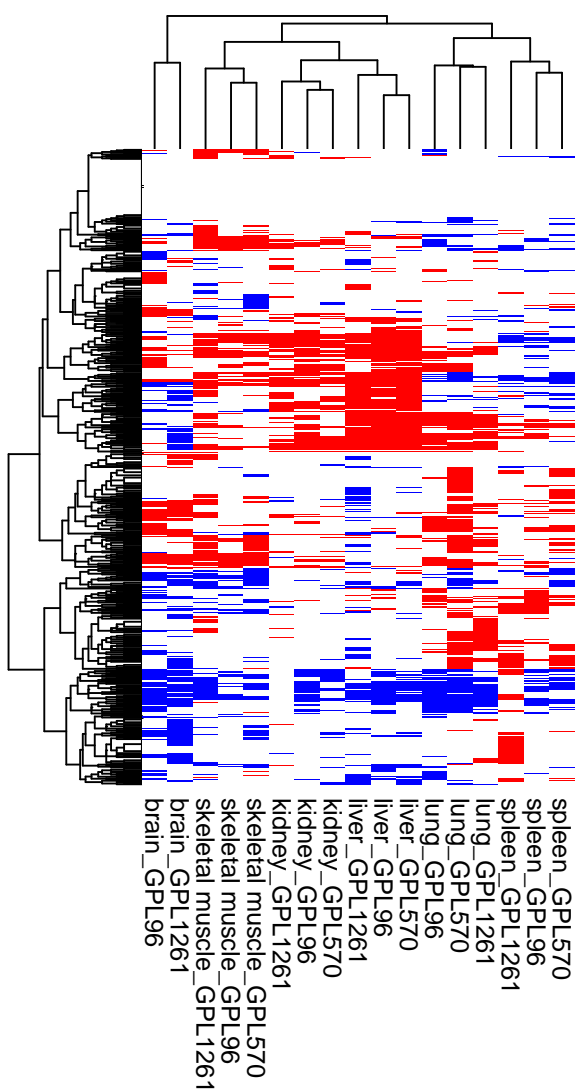


Figure 5: Heatmap formatted for the manuscript supplementary figure 4 - all pathways